GALE Go/No-Go Translation Evaluation Plan for the Summer 2006 Evaluation (GALE-06)

1 INTRODUCTION

The goal of the Global Autonomous Language Exploitation program (GALE) is to develop and apply computer software technologies to absorb, analyze and interpret huge volumes of speech and text in multiple languages. NIST is tasked with evaluating the "translation" aspect of GALE.

Specifically, the GALE Translation evaluation will test machine translation of text and machine transcription of audio, where transcription is defined as creating English text of foreign language audio. The test will include language data from both Arabic and Chinese, with system performance tallied separately for each language and separately for text and audio sources.

GALE contractors will be the only participants in this evaluation, and the participants must meet specific Go/No-Go levels of performance.

The first formal GALE Go/No-Go evaluation will be conducted in the summer of 2006. To prepare for GALE-06, NIST implemented a GALE Translation dry-run evaluation, with the focus on exercising the evaluation pipeline.

This document describes the evaluation protocols for this summer's GALE Translation evaluation.

2 EVALUATION TASKS

There are two tasks being evaluated in GALE-06, namely ***Translation*** and ***Transcription***. These tasks are being evaluated for two source languages, Arabic and Chinese.

2.1 TRANSLATION

Translation tests a system's ability to translate foreign text data into understandable and accurate English text. The input for this task will be a variety of mostly unstructured source language documents taken from newswire publications and web-base newsgroups. Systems must produce English text that completely captures the meaning conveyed by the source data, using easily understandable English.

2.2 TRANSCRIPTION

Transcription tests a system's ability to transcribe foreign language audio into understandable and accurate English text. The input for this task will be a variety of audio broadcasts from the news domain and from call-in talk shows. System must produce English text that completely captures the meaning conveyed by the source data, using easily understandable English.

3 EVALUATION DATA

The input data for the GALE Translation evaluation will consist of Arabic and Chinese language data from a variety of audio and text sources.

The evaluation data set provided to test GALE MT, may not be used to update other system components being evaluated separately for the GALE-06 Go/No-Go Evaluations.

All GALE-06 Translation evaluation data will be drawn from February 2006.

GALE teams may not use data that was published after February 1st, 2006 to train/build or test their system. Furthermore, any data that cannot be clearly identified as being produced before February 1st, 2006, is not to be used.

As agreed during the May-25th data committee conference call, any release of data, by the LDC, on or after June 15, 2006 may NOT be used for system development, tuning or testing in preparation of the GALE-06 Go/No-Go Translation evaluation.

3.1 TEXT SOURCES

Text data will come from a variety of "*Newswire*" and "*Web Newsgroup*" sources. The *newswire* sources will be very similar to recent DARPA Machine Translation (MT) evaluation test material. The web newsgroup data will be drawn from web logs and discussion forums which will include data that is less well formulated text.

There will be approximately 20,000 words from the text sources for each language under test, equally divided between *newswire* and *newsgroup* sources. Document length may range from 500-1500 words (as measured by the English reference).

The system under test will not have access to the categorization of each test document, that is, as to whether it is a newswire document or newsgroup data. Metadata that will be provided along with the text data include the date and the time (for newsgroup data the date and time refer to the time of the last post). This information will be included in the test document and identified by SGML attribute tags. All information included inside each source text document is available information to the system under test.

3.2 AUDIO SOURCES

Audio data will come from a variety of "*Broadcast News*" and "*Broadcast Conversations, or Talk Show*" sources. The *broadcast news* sources will be very similar to recent DARPA Speech-To-Text (STT) and Rich-Transcription (RT) evaluation test material. Broadcast conversation sources will focused more on round-table discussions and call-ins that have a conversational style of speech.

There will be approximately 3 hours of Arabic audio material and 2 hours of Chinese audio material[1], each divided equally between *broadcast news* and *broadcast conversation* sources.

[1] These are estimate durations. The goal is to create a test set with approximately the same amount of text data as audio data (as measured by the English reference text data). It is estimated that Arabic broadcast news contains about ~6000 words per hour, and Chinese broadcast news contains ~10,000 words per hour (1.5 characters = 1 word).

For both types of data, the test data will be excerpts of a longer broadcast.

The system under test will not have access to the categorization of each test segment, that is, as to whether it is broadcast news or broadcast conversation data. Metadata that will be provided along with the audio data include the source channel (ABC, CNN, …) and the date and the time of the recording. ~~This metadata will be provided in the form of a text table.~~

4 DATA FORMATS

The test data formats will be similar to the formats used in previous NIST evaluations, although the MT system translation output format will be required to include more information then was needed for previous NIST MT evaluations.

4.1 INPUT FORMATS

This section describes the formats of the source data files that will be distributed as evaluation test data for use in the GALE-06 Translation evaluation.

4.1.1 Text Input

The source data for text input will be UTF-8 encoded files, with each file corresponding to a single document. The native text inside each document will not be pre-segmented as it was for previous NIST MT evaluations.

Each document will contain a series of SGML tags that are used for document identification and document structure. Only the native source language that is not embedded in a SGML tag is to be translated. *(Newsgroup data may have a ”<QUOTE” tag that contains native source text as an attribute. This data is not to be translated).*

Commonly occurring SGML tags that surround text to be translated include: beginning and ending (**headline**), (**text**), (**post**), and/or (**subject**) tags.

NOTE: Text between beginning and ending (**poster**) tags should not be translated. They often contain “handles” which are not necessarily meaningful and sometimes they are completely incoherent.

An example of a source file for text input:

```
  <doc id=”NYT-doc1”>
  <body>
  <headline> ARABIC LANGUAGE TEXT
  </headline>
  <text>
  ARABIC LANGUAGE TEXT

  ARABIC LANGUAGE TEXT
  …
</text>
</body>
</doc>
```

4.1.2 Audio Input

The source data for audio input will be a separate audio waveform for each broadcast. Each waveform will be a complete story segment. They will not contain commercials. Each waveform will be distributed in 16-bit PCM format and will include a NIST SPHERE header.

There will be one Un-partitioned Evaluations Maps (UEM) file which specifies the time regions within each audio recording to be evaluated.

The UEM file structure is as follows:

<F><SP><C><SP><BT><SP><ET>

where

<F> indicates the file id, consisting of the path, filename and extension of the waveform to be processed.

<SP> indicates a space (“ “).

<C> indicates the waveform channel, which, for GALE-06 is always set to 1.

<BT> indicates the beginning time of the segment measured in seconds from the beginning of the file which is time 0.

<ET> indicates the ending time of the segment measured in seconds from the beginning of the file which is time 0.

~~There will be one text file that identifies the allowable metadata for each broadcast. This file will be named “metadata.txt”~~

4.2 OUTPUT FORMATS

This section describes the file formats that the systems evaluated for GALE-06 Translation, must produce.

Each GALE team is to submit system translations for ***exactly*** one system. *Contrastive* systems will not be evaluated. Translations must be submitted for the complete evaluation test set for each language.

System translations should have proper capitalization and all punctuation should be attached to the text.

4.2.1 MT Output from Text

The system translations of text sources must adhere to the following NIST MT data format.

Each system translation must preserve all SGML tags present in the original source document. Many of these tags carry document identification information. Although the original source text documents do not contain segment information, the system translation output will be required to include segmentation. This will be accomplished by adding a series of segments tags around the translated text. Each segment tag must have an id attribute which sequentially identifies the segments. Each segment tag has a corresponding closing tag. An example of a MT text translation file:

```
<doc id=”NYT-doc1”>
<body>
<headline>
<seg id=”1”> TRANSLATED ENGLISH TEXT </seg>
</headline>
<text>
<seg id=”2”> TRANSLATED ENGLISH TEXT </seg>
<seg id=”3”> TRANSLATED ENGLISH TEXT </seg>
<seg id=”4”> TRANSLATED ENGLISH TEXT </seg>
…
</text>
</body>
</doc>
```

It will be the system’s responsibility to provide segmented output. A segment should be defined to be a sentence, or a sentence-like unit. *Note: excessively longer or extremely shorter segmentation **might** put an undue burden on the post editors.*

4.2.2 MT Output from Audio

The MT system translation output format from audio is exactly the same as for text, but **must** contain the optional segment attributes for time boundaries (start & end):

```
<audiofile fileid="CNN_HEADLINE-file1">
<seg id="1" start="1.25" end="12.33">
TRANSLATED ENGLISH TEXT </seg>
<seg id="2" start="20.95" end="55.42">
TRANSLATED ENGLISH TEXT </seg>
…
</audiofile>
```

5 Reference Data

System translation output will be evaluated (post-edited) by comparing the system output with a single gold-standard English reference translation. NVTC will be responsible for creating the GALE Translation reference data.

In cases where the original source language is ambiguous, the reference data will contain allowable alternatives for words or phrases.

6 Data Preprocessing

This section describes the preprocessing of reference and system translation data that will occur *before* the system translations are evaluated.

6.1 MT Alignment to Reference

MT systems and human translators will not always agree on sentence boundaries and/or sentence like units for the translations of the foreign source data

System MT translations will be aligned with the reference translations before being post edited for evaluation. This alignment is necessary for the post-editing process described in section 7.1. The better the alignment, the less of a burden it will be to the post editors, but perfect alignment is not necessary (post editors are instructed/trained to look for equivalent meaning ahead and behind the current segment they are editing).

6.1.1 Alignment of Text

System and reference translations of the text data will be aligned to maximize the likelihood that the segments under focus contain the translations of the same source sentence.

> *Note*: The post editor tool displays the complete reference translation in a column of three rows. The first row contains everything that corresponds to what has already been edited, the middle row is to match what is currently being edited, and the third row contains the translations corresponding to the remainder of the document. A second column of three rows list the corresponding system translation.

For the GALE evaluation, NIST will automatically align the system translations to the reference translations using Aachen's "mwerSegmenter", which back traces the decisions of the Levenshtein edit distance algorithm, to match system translation to the reference translation's segmentation points. *This is a different technique than was used for the GALE Translation dry run.*

All aligned data will be hand checked by NIST before going to the post editor. Perfect alignment isn't necessary, but NIST will insure the system and reference translations do not become grossly out of sync.

6.1.2 Alignment of Audio

System and reference translations of audio data will be aligned to maximize the likelihood that the segments under focus contain the translations of the same sentence. The alignment will be produced using the same software as is used for the text sources, and all alignments will be hand checked before they go to the post editor.

The time stamps will be used to resolve questions of apparent poor alignments. It is not necessary for the system translation to match the reference translation exactly because the post-editing is to be viewed as taking place on the document as a whole. (The alignment is an effort to simplify the information load on the post editors.)

7 Evaluation Metric

GALE-06 will use an edit-distance metric to evaluate system translation quality. This will be accomplished by having one or more qualified human editor(s)[2] make changes to the MT output so that the resulting edited-MT output uses understandable English that contains exactly the same information as the reference data in as few edits as they can use. The editors will be given specific guidelines[3] to follow while performing the edits.

7.1 Post Editing Process

NIST has developed an editing interface[4] designed for the post editing task. An editor will view the contents of one complete document[5] with the focus on a single sentence-like unit. The reference file will be aligned with the system translation and displayed in two separate columns. Alternative words and phrases will be given to the editor in instances when the original source language data was ambiguous or if independent translators did not agree on the exact meaning.

The post editor will modify the segment under focus until they feel that the MT output completely captures the meaning conveyed in the reference data, and nothing more. They are instructed to make modifications using as few edits as possible. Although the editor will be looking at the aligned segments, they will be free to use context before and after the current line of focus. See the post editing guidelines for more details.

Each translated document, by each system, will be post-edited by 3 editors. Selected edited documents will be reviewed in a second pass. There will be quality control measures in place to verify that the post editors are performing their job in an acceptable manner. The exact amount of data that will be reviewed will be affected by how fast the editors are getting through the test data.

7.2 The Edit Distance Metric

NIST will compare the resulting edited-MT with the original MT and count the number of edits. Each edit is weighted equally. The number reported will be the ratio of the number of edits to the number of words in the gold standard reference data. In the case of alternative words and phrases, only the first choice listed will be counted as part of the reference.

[2] For GALE-06 NIST will contract with the LDC to hire qualified post editors and implement the post editing.

[3] The "Post Editing Guidelines For GALE Machine Translation Evaluation" maybe accessed via the NIST GALE website at: URL https://www.nist.gov/speech/tests/gale/2006/doc/

[4] The JAVA based post editing interface maybe accessed via the NIST GALE website at: https://www.nist.gov/speech/tests/gale/2006/software/

[5] A document could be a complete newswire document, a broadcast news story, a section taken from a user group, or a section from a broadcast talk show.

This score will be automatically calculated using the BBN supplied evaluation script calc_ter_v5.pl[6]. NIST will report the mean and low HTER scores over the first-pass edited data. NIST will report the mean and low HTER scores over data that has received a second pass. This second pass set will be more reflective of the true edit distance.

NIST will distribute to the GALE community, all of the alignments and HTER scores for each document and for each data genre. Scores will be distributed on the document and segment level.

8 Submitting Results to NIST

E-mail[7] is the preferred method for submitting system translation files to NIST.

8.1 Machine Translation Output

NIST will score one set of machine translations from each participant using the above mentioned post-editing protocols. *Contrastive systems* will not be evaluated.

8.2 Packaging System Translations

Create a directory that identifies the GALE team.

“Agile”, “Rosetta”, or “Nightingale”

Under your team directory create the following structure:’

./Arabic/audio
./Arabic/text
./Chinese/audio
./Chinese/text

Place the system translations in their proper directory.

System translation files should have the same name as the input file but replace audio file “.sph” and text file “.sgm” extensions with “.system.sgm”

8.3 System Descriptions

A system description will NOT be required for the GALE evaluation.

9 Schedule

Date	Event
Feb-01-2006	Begin collection of GALE evaluation data.
Feb-28-2006	End collection of GALE evaluation data.
Jun-22-2006	GALE Translation evaluation begins.
Jul-06-2006	Translations of TEXT data, due at NIST
Jul-11-2006	Post Editing begins.
Jul-13-2006	Translations of AUDIO data, due at NIST
Aug-31-2006	Post Editing ends.
TBD	GALE Evaluation and PI meeting

[6] The BBN supplied evaluation script is available via the NIST GALE website at: https://www.nist.gov/speech/tests/gale/2006/software

[7] GALE-06 machine translation output should be e-mailed to gale_poc@nist.gov.